ICFHR 2014 COMPETITION ON HANDWRITTEN KEYWORD SPOTTING (H-KWS 2014)

IOANNIS PRATIKAKIS[1]

KONSTANTINOS ZAGORIS[1,2]

BASILIS GATOS[2]

GEORGIOS LOULOUDIS[2]

NIKOLAOS STAMATOPOULOS[2]

[1] Visual Computing Group
Democritus University of Thrace
Dept. of Electrical and Computer Engineering
Xanthi, Greece

[2] National Centre of Scientific Research "Demokritos"
Institute of Informatics and Telecommunications
Athens, Greece

INTRODUCING ICFHR 2014 H-KWS

Handwritten keyword spotting is the task of detecting query words in handwritten document image collections without explicitly recognizing it.

The objective of the H-KWS 2014 is threefold:

- Record **current advances** in keyword spotting.
- Provide **benchmarking handwritten datasets** containing both historical and modern documents from multiple writers.
- Explore **established evaluation performance measures** frequently encountered in the information retrieval literature while providing the **software** for these measures as **implementation reference**.

ORGANIZING ICFHR 2014 H-KWS

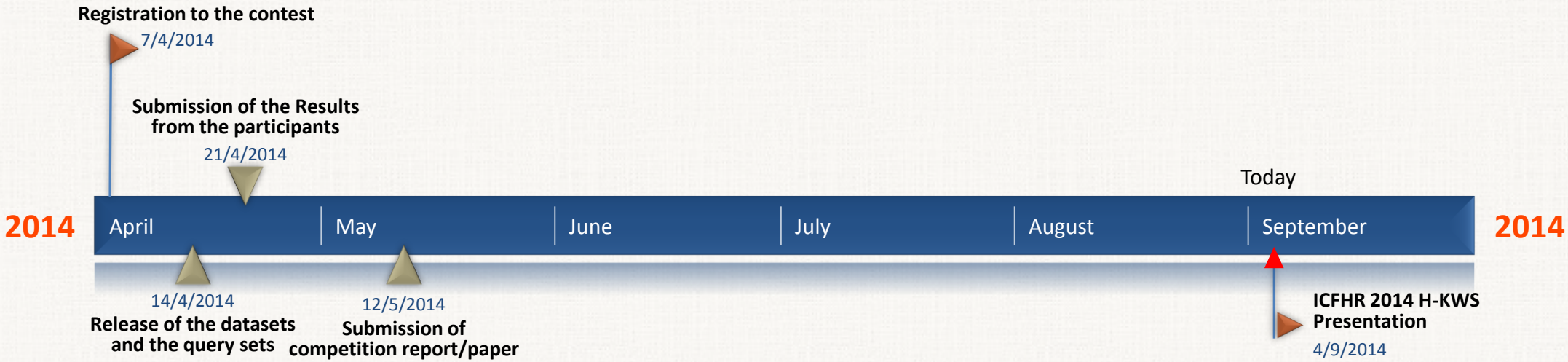
TRACK I - SEGMENTATION-BASED

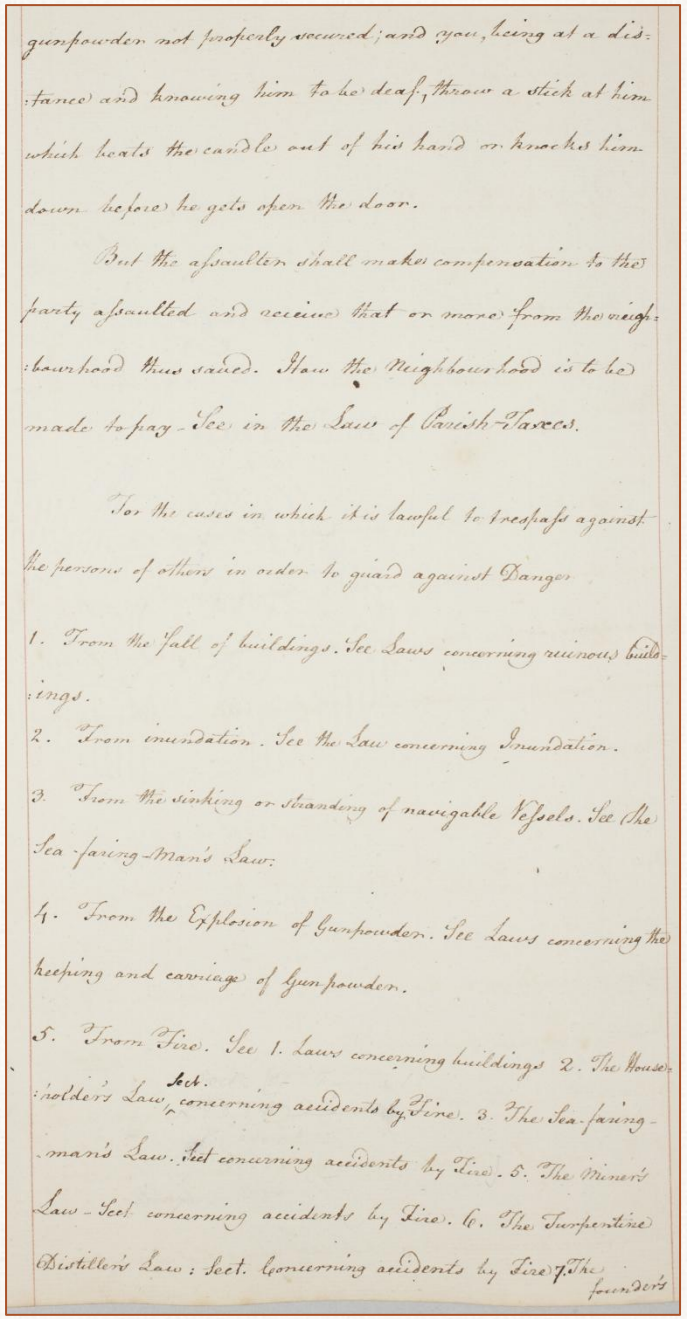
- 50 document images of Bentham dataset.
- 100 document images of Modern dataset (25 documents per language).
- Word Location in XML format.

TRACK II - SEGMENTATION-FREE

- 50 document images of Bentham dataset.
- 100 document images of Modern dataset (25 documents per language).

ICFHR 2014 H-KWS TIMELINE



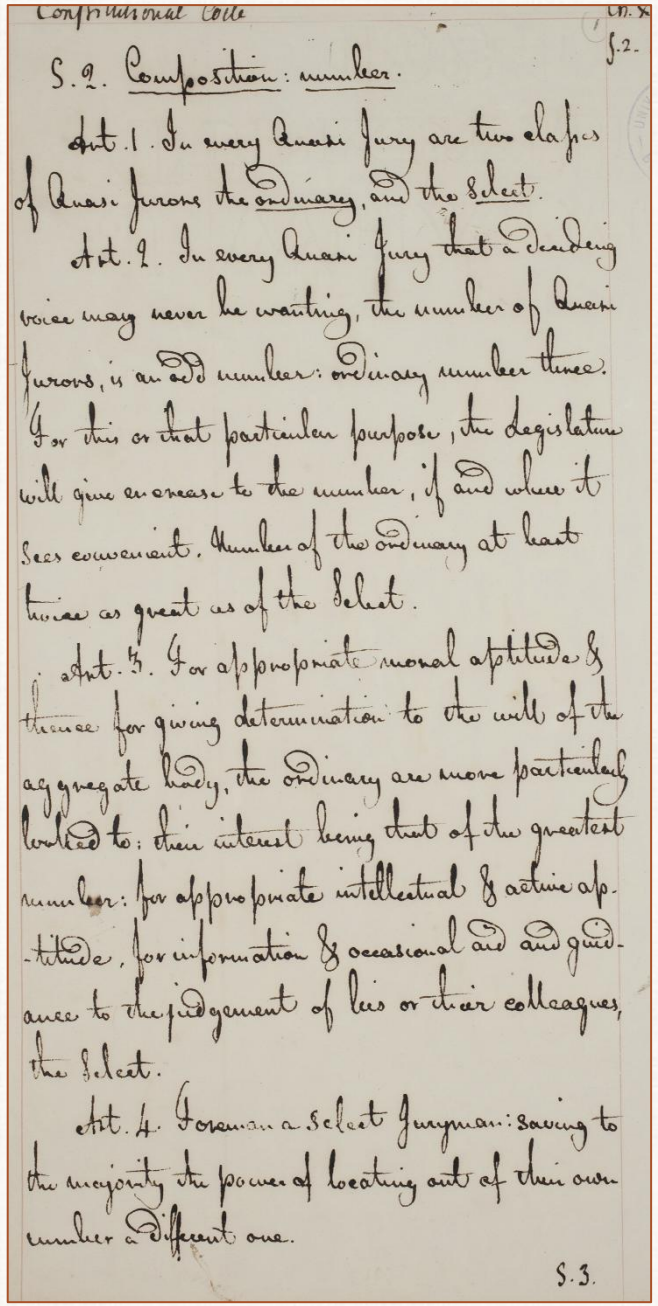


gunpowder not properly secured; and you, being at a distance and knowing him to be deaf, throw a stick at him which beats the candle out of his hand or knocks him down before he gets open the door.

But the assaulter shall make compensation to the party assaulted and accine that or more from the neighbourhood thus saved. How the Neighbourhood is to be made to pay - See in the Laws of Parish Taxes.

For the cases in which it is lawful to trespass against the persons of others in order to guard against Danger

1. From the fall of buildings. See Laws concerning ruinous buildings.
2. From inundation. See the Law concerning Inundation.
3. From the sinking or stranding of navigable Vessels. See the Sea-faring-Man's Law.
4. From the Explosion of Gunpowder. See Laws concerning the keeping and carriage of Gunpowder.
5. From Fire. See 1. Laws concerning buildings 2. The Householder's Law, Sect. concerning accidents by Fire. 3. The Sea-faring-man's Law. Sect concerning accidents by Fire. 5. The Miner's Law - Sect concerning accidents by Fire. 6. The Turpentine Distiller's Law: Sect. Concerning accidents by Fire 7. The founder's



Constitutional Code Ch. X
S.2.

S.2. Composition: number.

Art. 1. In every Quasi Jury are two classes of Quasi jurors the ordinary, and the Select.

Art. 2. In every Quasi Jury that a deciding voice may never be wanting, the number of Quasi Jurors, is an odd number: ordinary number three. For this or that particular purpose, the legislature will give an increase to the number, if and where it sees convenient. Number of the ordinary at least twice as great as of the Select.

Art. 3. For appropriate moral aptitude & fitness for giving determination to the will of the aggregate body, the ordinary are more particularly looked to: their interest being that of the greatest number: for appropriate intellectual & active aptitude, for information & occasional aid and guidance to the judgement of his or their colleagues, the Select.

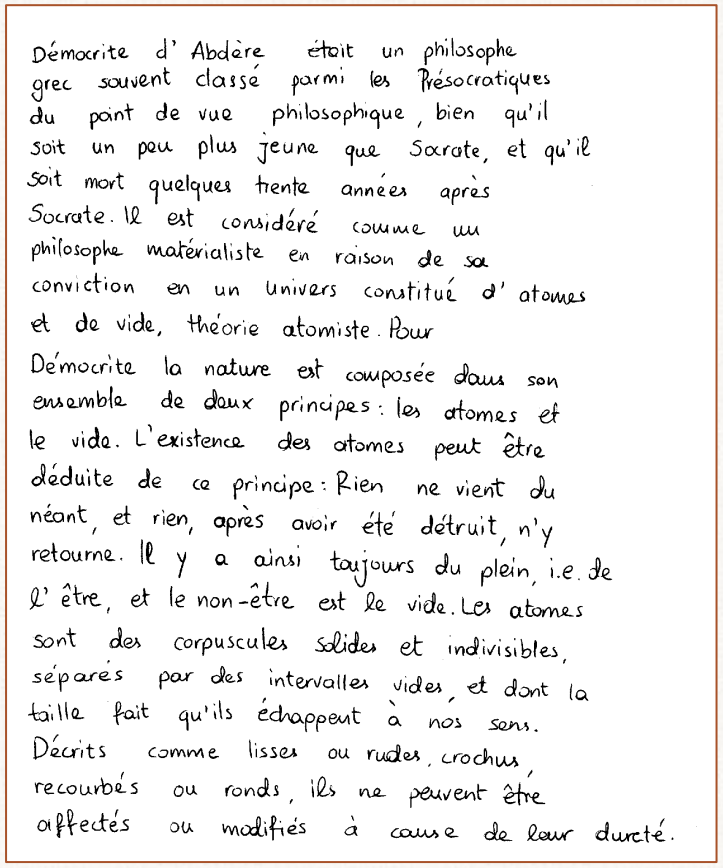
Art. 4. Foreman a Select Juryman: saving to the majority the power of locating out of their own number a different one.

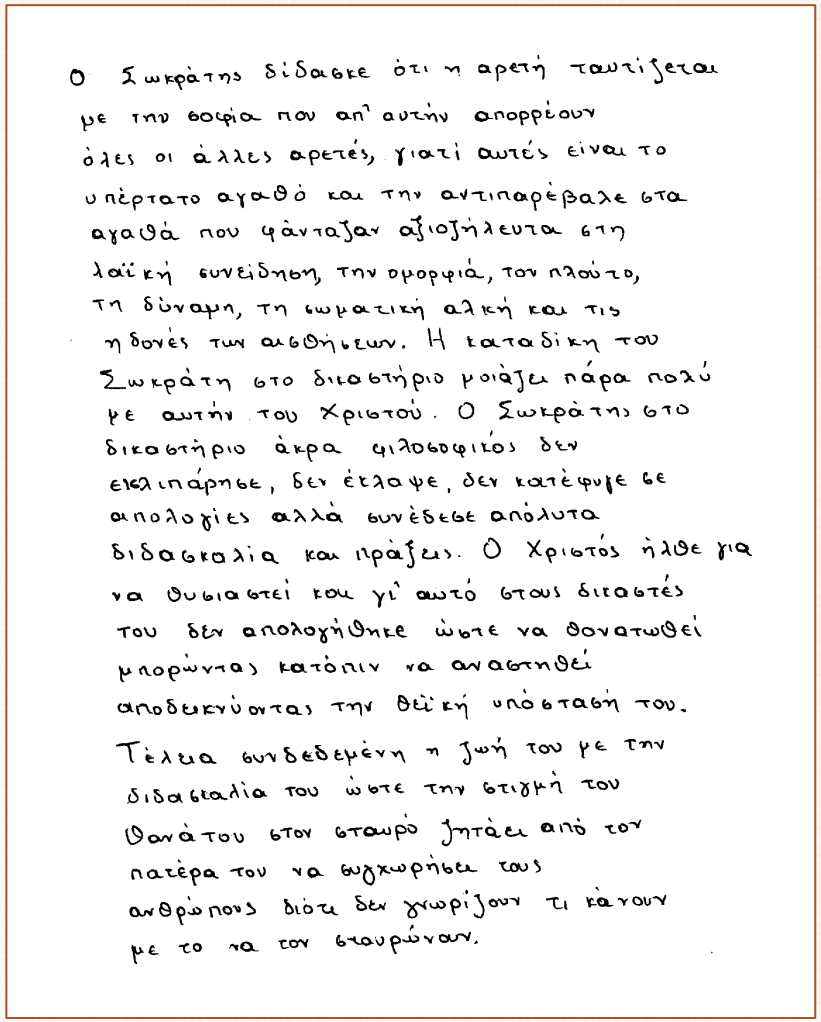
S.3.

BENTHAM DATASET

It consists of high quality (approximately 3000 pixel width and 4000 pixel height) handwritten manuscripts.

The documents are written by Jeremy Bentham (1748-1832) himself as well as by Bentham's secretarial staff over a period of sixty years.


Démocrite d' Abdère était un philosophe grec souvent classé parmi les Présocratiques du point de vue philosophique, bien qu'il soit un peu plus jeune que Socrate, et qu'il soit mort quelques trente années après Socrate. Il est considéré comme un philosophe matérialiste en raison de sa conviction en un Univers constitué d' atomes et de vide, théorie atomiste. Pour Démocrite la nature est composée dans son ensemble de deux principes: les atomes et le vide. L'existence des atomes peut être déduite de ce principe: Rien ne vient du néant, et rien, après avoir été détruit, n'y retourne. Il y a ainsi toujours du plein, i.e. de l' être, et le non-être est le vide. Les atomes sont des corpuscules solides et indivisibles, séparés par des intervalles vides, et dont la taille fait qu'ils échappent à nos sens. Décrits comme lisses ou rudes, crochus, recourbés ou ronds, ils ne peuvent être affectés ou modifiés à cause de leur dureté.


Ο Σωκράτης δίδαξε ότι η αρετή ταυτίζεται με την σοφία που απ' αυτήν απορρέουν όλες οι άλλες αρετές, γιατί αυτές είναι το υπέρτατο αγαθό και την αντιπαρέβαλε στα αγαθά που φάνταζαν αξιοζήλευτα στη λαϊκή συνείδηση, την ομορφιά, τον πλούτο, τη δύναμη, τη σωματική αλκή και τις ηδονές των αισθήσεων. Η καταδίκη του Σωκράτη στο δικαστήριο μοιάζει πάρα πολύ με αυτήν του Χριστού. Ο Σωκράτης στο δικαστήριο άκρα φιλοσοφικός δεν εκλιπάρησε, δεν έκλαψε, δεν κατέφυγε σε απολογίες αλλά συνέδεσε απόλυτα διδασκαλία και πράξεις. Ο Χριστός ήλθε για να θυσιαστεί και γι' αυτό στους δικαστές του δεν απολογήθηκε ώστε να θανατωθεί μπορώντας κατόπιν να αναστηθεί αποδεικνύοντας την θεϊκή υπόστασή του. Τέλεια συνδεδεμένη η ζωή του με την διδασκαλία του ώστε την στιγμή του θανάτου στον σταυρό ζητάει από τον πατέρα του να συγχωρήσει τους ανθρώπους διότι δεν χωρίζουν τι κάνουν με το να τον σταυρώνουν.

MODERN DATASET

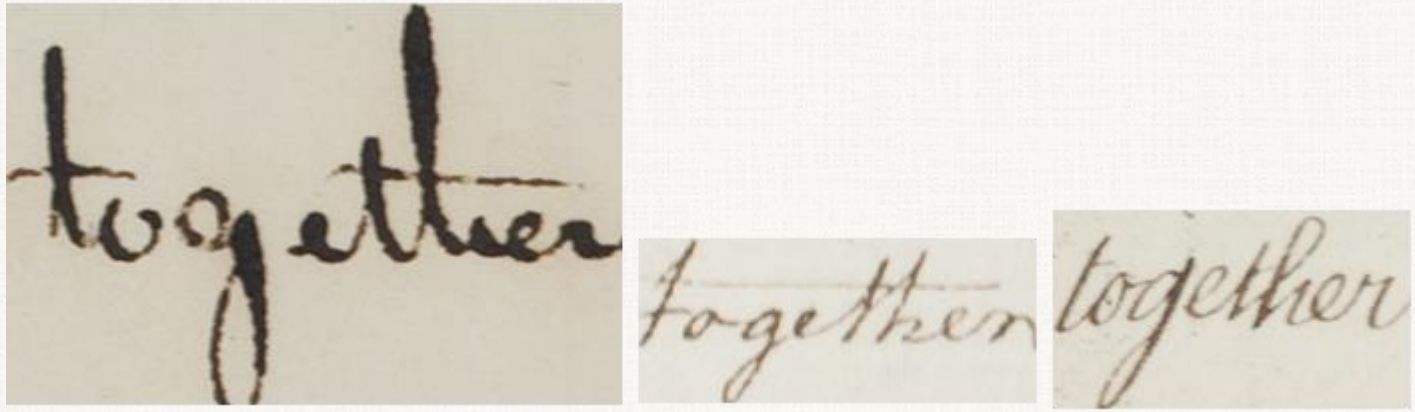
It consists of **modern handwritten documents** from the ICDAR 2009 Handwritten Segmentation Contest.

These documents originate from **several writers** that were asked to copy a given text.

They do not include any non-text elements (lines, drawings, etc.).

They are written in four (4) languages: **English**, **French**, **German** and **Greek.**

BENTHAM



MODERN

CHALLENGES

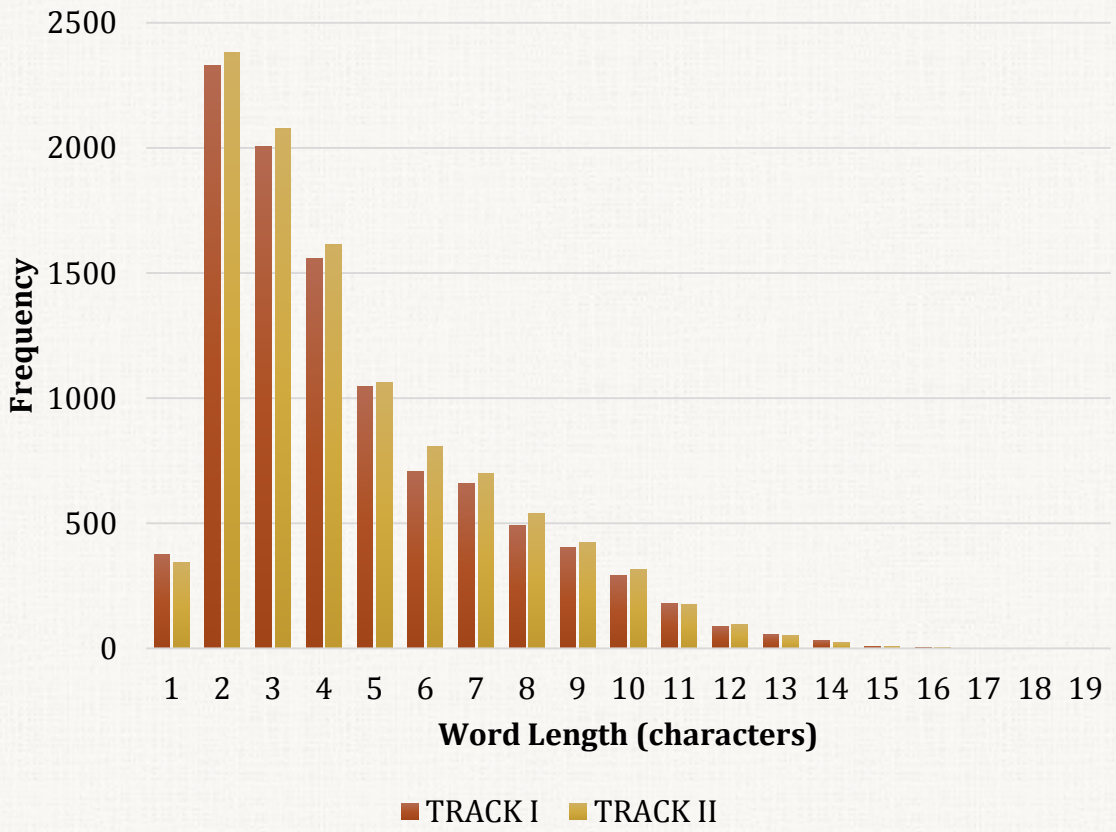
They both contain several very difficult problems to be addressed, wherein the most difficult is the word variability.

The variation of the same word is high and involves:

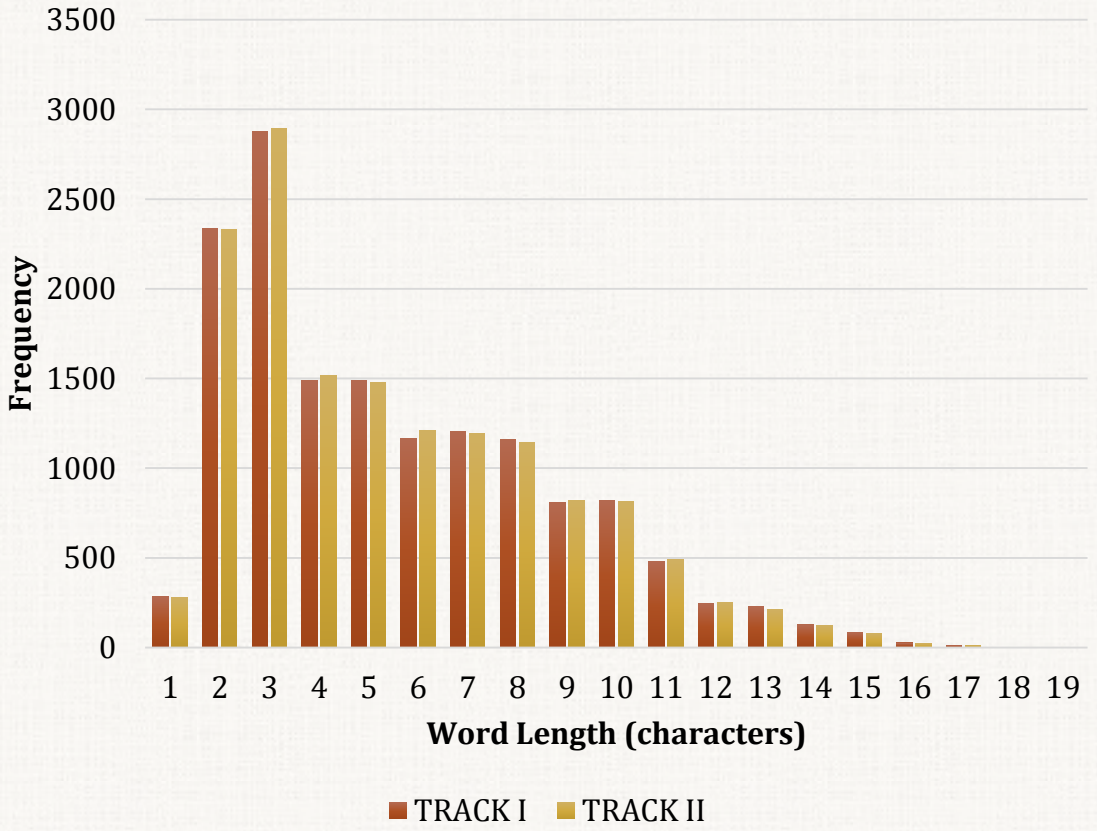
- writing style
- font size
- noise
- their combination

WORD-LENGTH STATISTICS FOR EACH DATASET

BENTHAM



MODERN



QUERY STATISTICS

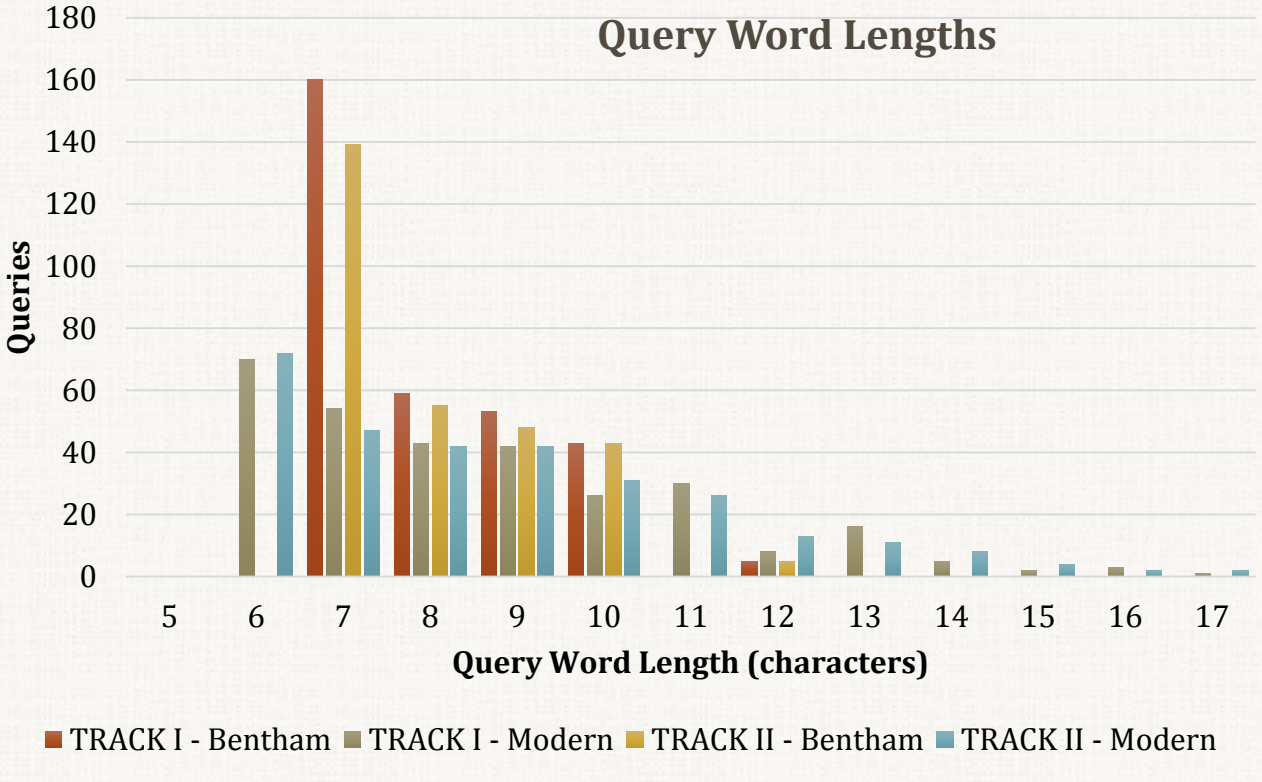
The query set of each dataset is provided in XML format and it contains word image queries of length greater than 6 and frequency greater than 5.

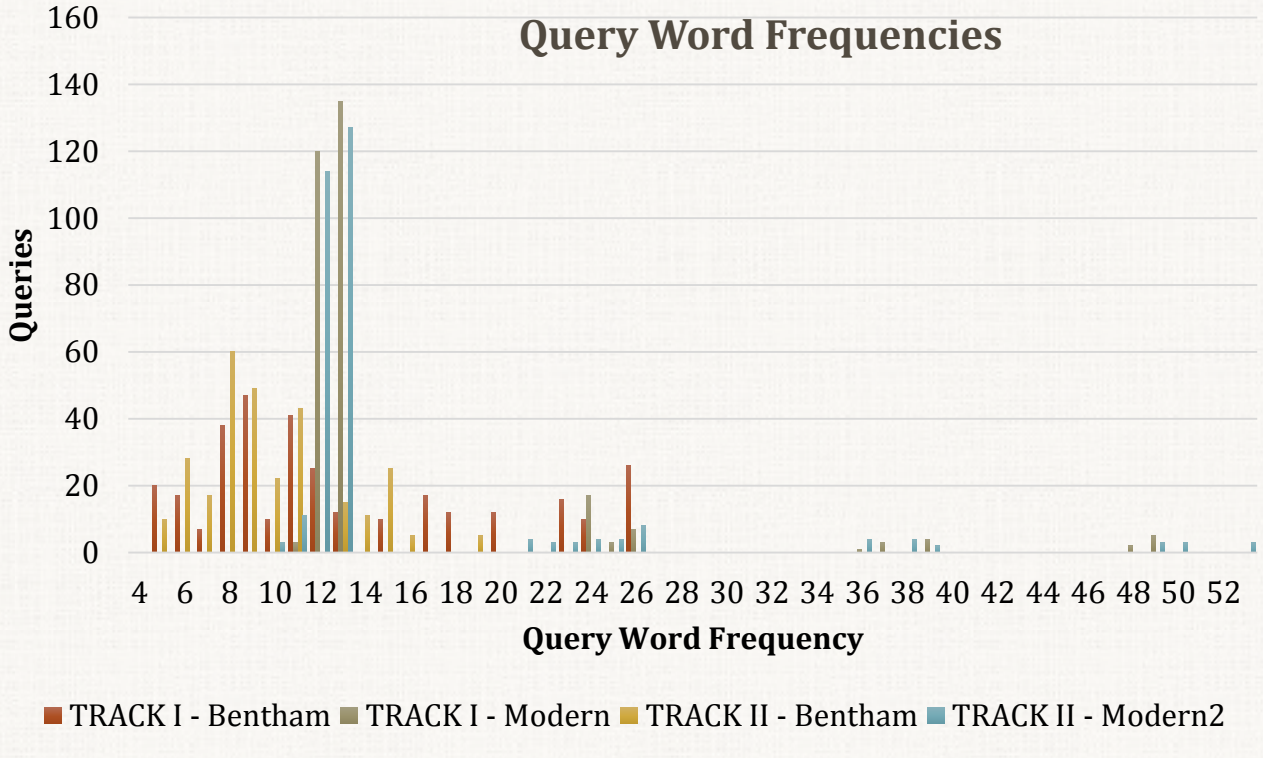
TRACK I - SEGMENTATION-BASED

- 320 queries for the Bentham Database
- 300 queries for the Modern Database

TRACK II - SEGMENTATION-FREE

- 290 queries for the Bentham Database
- 300 queries for the Modern Database





http://vc.ee.duth.gr/H-KWS2014/#Datasets

Datasets

TRACK I:SEGMENTATION-BASED

Bentham

Dataset

ICFHR 2014 Query Set

Modern

Dataset

ICFHR 2014 Query Set

TRACK II:SEGMENTATION-FREE

Bentham

Dataset

ICFHR 2014 Query Set

Modern

Dataset

ICFHR 2014 Query Set

EVALUATION CHALLENGES

- Small variations of the query word that can be found in the datasets. For example the word “husband” appears as well as :
  - husband**,**
  - husband**:**
  - husband**.**
  - **H**usband**.**
  - **H**usband**]**
- Evaluating overall performance as well as precision.
- Segmentation - free systems may not detect the whole word or include parts of another word.

CHOSEN EVALUATION MEASURES

- **Precision at Top 5 Retrieved words** (**P@5**) for evaluating precision performance.
- The **Mean Average Precision** (**MAP**) for evaluating overall performance.
- **Normalized Discounted Cumulative Gain** (**NDCG**) with **binary** judgment relevancies for evaluating precision-oriented overall performance.
- **Normalized Discounted Cumulative Gain** (**NDCG**) with **non-binary** judgment relevancies for evaluating small variations of the query word.

PRECISION AT TOP K RETRIEVED WORD (P@K)

Precision is the fraction of retrieved words that are relevant to the query.

P@k is the precision for the **k top** retrieved **words**.

In the proposed evaluation, **P@5** is used which is the precision at **top 5** retrieved **words**.

This metric defines how successfully the algorithms produce relevant results to the first 5 positions of the ranking list.

- $$P@k = \frac{|\{relevant\ words\} \cap \{k\ retrieved\ words\}|}{|\{k\ retrieved\ words\}|}$$

MEAN AVERAGE PRECISION (MAP)

It is a typical measure for the performance of information retrieval systems.

It is implemented from the Text REtrieval Conference (TREC) community by the National Institute of Standards and Technology (NIST).

It is defined as the average of the precision value obtained after each relevant word is retrieved:

$$AP = \frac{\sum_{k=1}^{n}\left(P@K \times rel(k)\right)}{\{relevant\ words\}} \qquad \text{where:} \qquad rel(k) = \begin{cases} 1, & if\ word\ at\ rank\ k\ is\ relevant \\ 0, & if\ word\ at\ rank\ k\ is\ not\ relevant \end{cases}$$

NORMALIZED DISCOUNTED CUMULATIVE GAIN (NDCG)

The NDCG measures the performance of a retrieval system based on the graded relevance of the retrieved entities.

It varies from 0.0 to 1.0, with 1.0 representing the ideal ranking of the entities.

It is defined as:

$$nDCG = \frac{DCG}{IDCG} \qquad \text{where:} \qquad DCG = rel_1 + \sum_{i=1}^{n} \frac{rel_i}{log_2(i+1)}$$

rel_i is the relevance judgment at position i and $IDCG$ is the ideal DCG which is computed from the perfect retrieval result.

NON-BINARY VS BINARY RELEVANCE JUDGMENT VALUES

NON-BINARY

Word	Relevance Judgment
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband,	0.9
husband:	0.9
husband.	0.9
Husband.	0.8
Husband]	0.8

BINARY

Word	Relevance Judgment
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband	1.0
husband,	1.0
husband:	1.0
husband.	1.0
Husband.	1.0
Husband]	1.0

SEGMENTATION – FREE OVERLAPPING THRESHOLD

- A word instance is considered as detected only if there is a **significant overlap** with the ground truth word.

The overlap is expressed by the **intersection** over the **ground truth word area** metric (*IOA*) and it is defined as:

$$IOA = \frac{A \cap B}{A}$$

where A and B denote the bounding box areas of the ground truth word and the method output word, respectively.

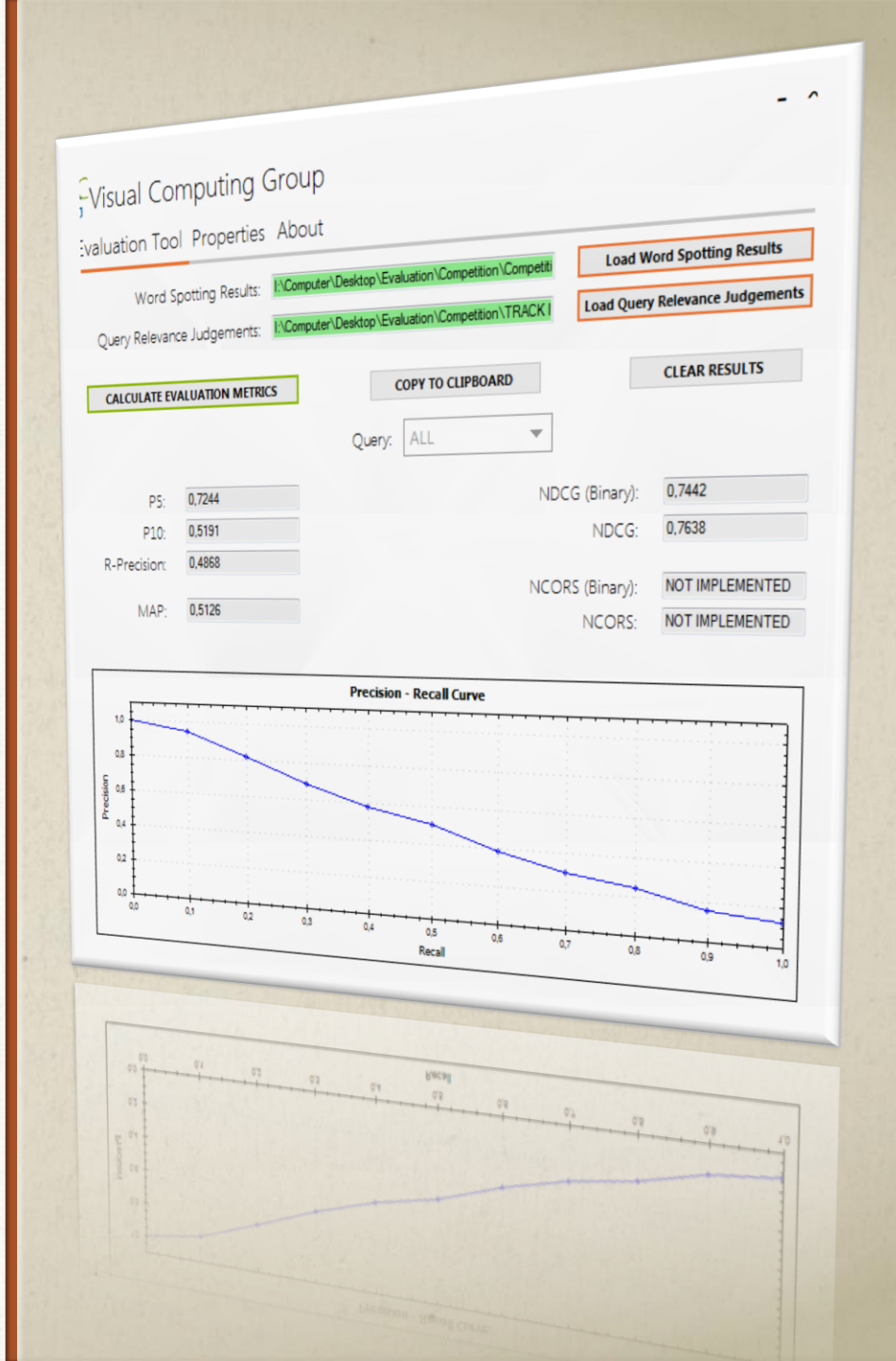
The *IOA* metric ranges from 0 to 1, where 1 corresponds to exact matching.

A **threshold** T is applied in order to decide whether the word instance and the segmented word match sufficiently.

The performance evaluation for three different thresholds (**0.6**, **0.7** and **0.8**) is used for testing.

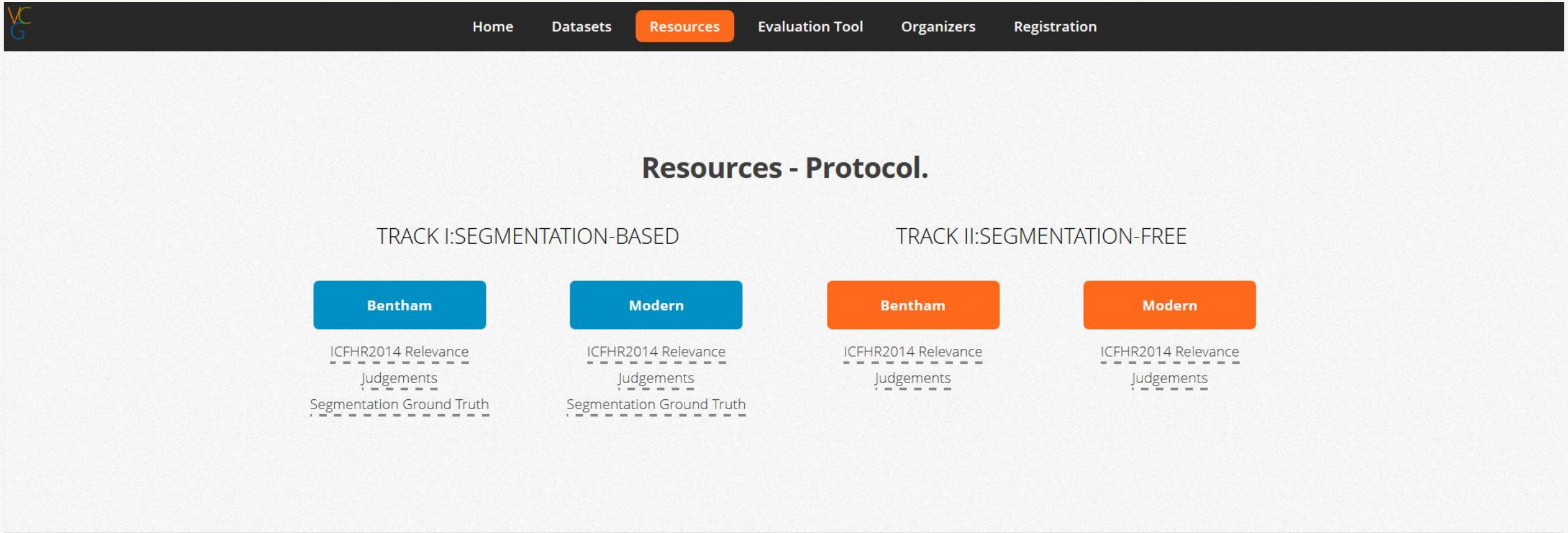
EVALUATION APPLICATION

- An evaluation application is developed as referenced implementation for each metric.
- It is available for **Windows**, **Mac OS X** and **Linux** operating systems as both **command-line** and **GUI form.**
- It accepts as input the **experimental results file** and the **relevance judgment file**, which represents the ground truth. Afterwards, it calculates the aforementioned evaluation metrics.



http://vc.ee.duth.gr/H-KWS2014/#VCGEval

http://vc.ee.duth.gr/H-KWS2014/#Resources



PARTICIPANTS

Method	Affiliation		Participating
G1	The Blavatnik School of Computer Science, Tel-Aviv University, Israel		TRACK I TRACK II
G2	Computer Vision Center, Barcelona, Spain		TRACK I
G3	Smith College Department of Computer Science, Northampton MA, USA		TRACK I TRACK II
G4	Université de Lyon, CNRS INSA - Lyon, LIRIS, France		TRACK II
G5	Institute for Communications Technology (IfN) of Technische Universität Braunschweig, Braunschweig, Germany		TRACK II

TRACK I: SEGMENTATION-BASED - EXPERIMENTAL RESULTS

BENTHAM DATASET

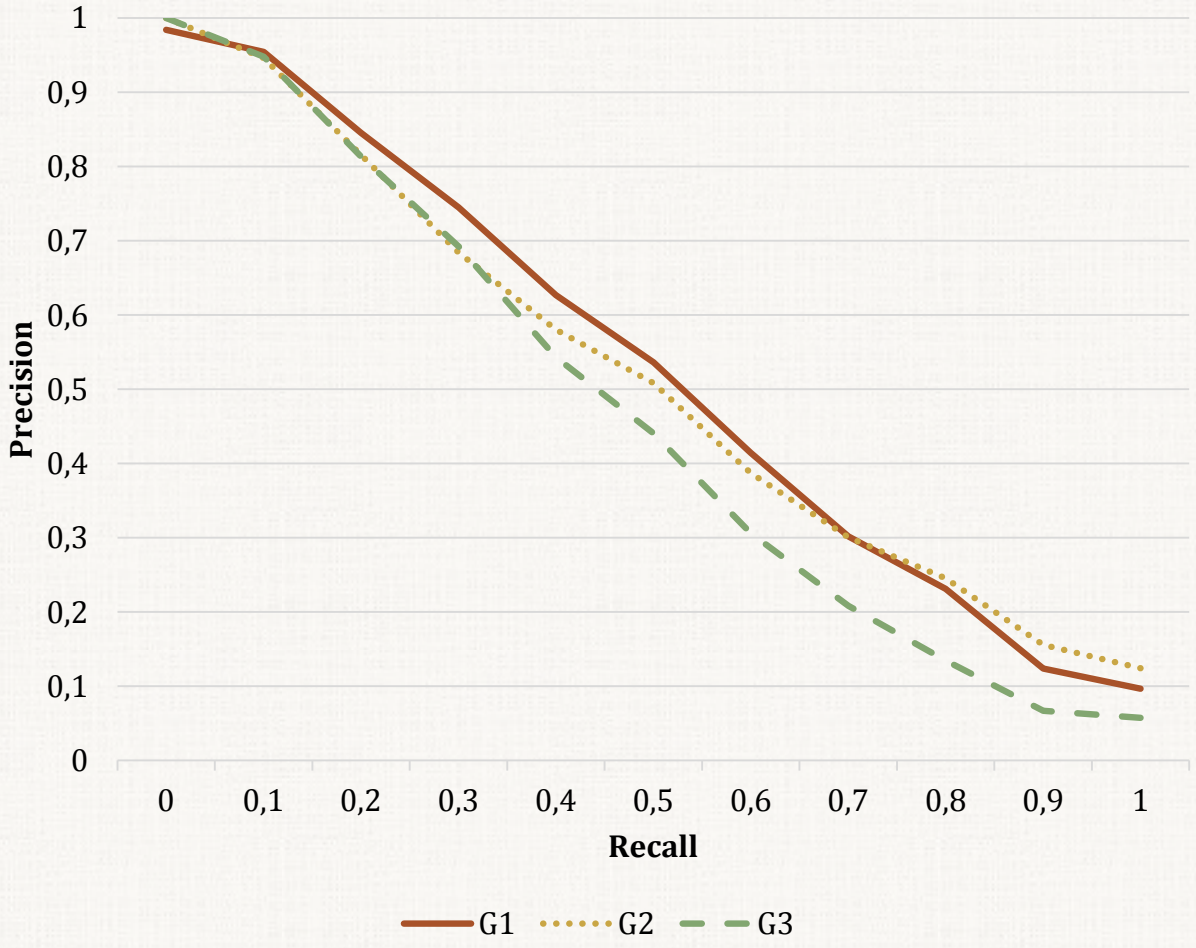
Method	P@5	MAP	NDCG(Binary)	NDCG
G1	0.738 (**1**)	0.524 (**1**)	0.742 (**2**)	0.762 (**2**)
G2	0.724 (**2**)	0.513 (**2**)	0.744 (**1**)	0.764 (**1**)
G3	0.718 (**3**)	0.462 (**3**)	0.638 (**3**)	0.657 (**3**)

MODERN DATASET

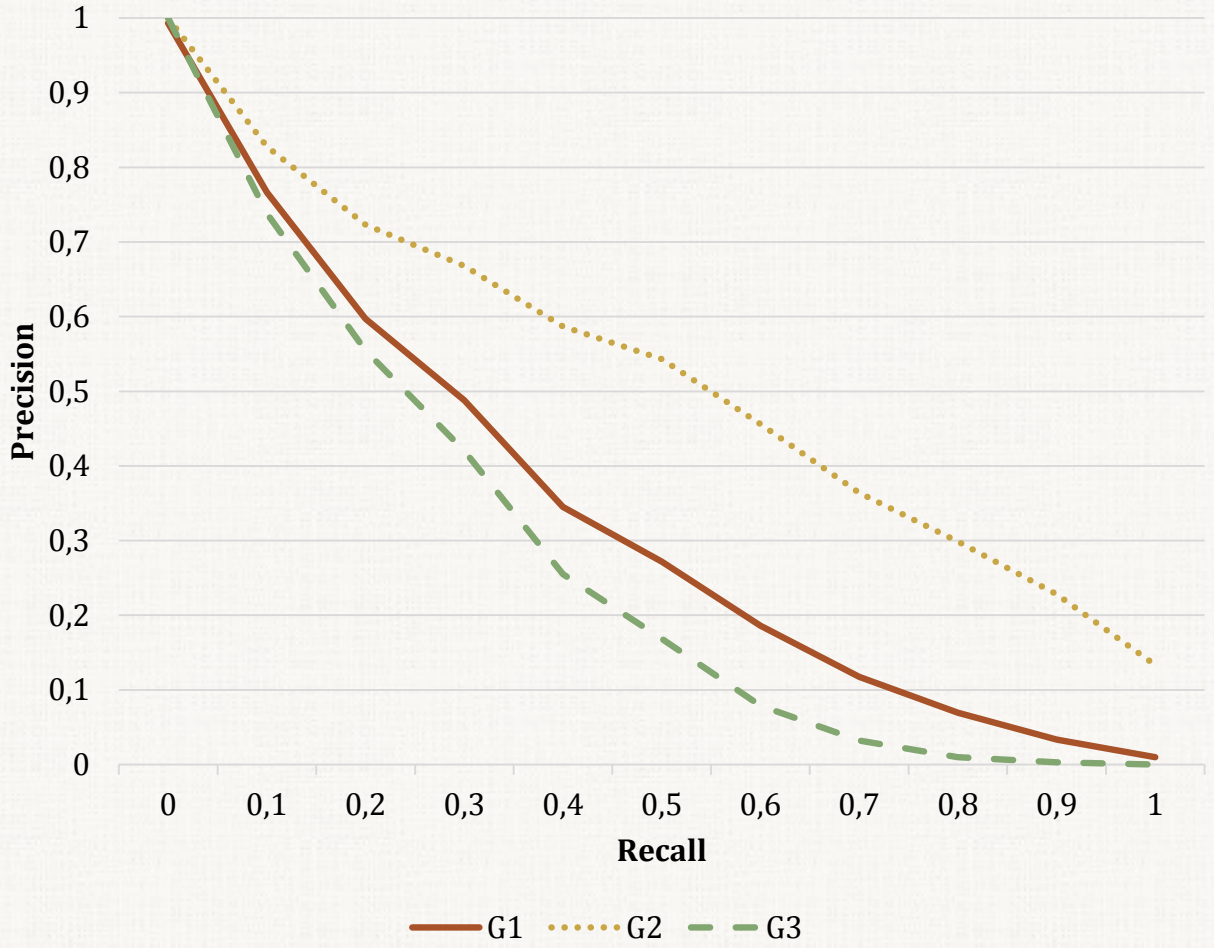
Method	P@5	MAP	NDCG(Binary)	NDCG
G1	0.588 (**2**)	0.338 (**2**)	0.611 (**2**)	0.612 (**2**)
G2	0.706 (**1**)	0.523 (**1**)	0.757 (**1**)	0.757 (**1**)
G3	0.569 (**3**)	0.278 (**3**)	0.484 (**3**)	0.484 (**3**)

TRACK I: SEGMENTATION-BASED - PRECISION – RECALL CURVES

BENTHAM DATASET



MODERN DATASET



TRACK II: SEGMENTATION-FREE - EXPERIMENTAL RESULTS

BENTHAM DATASET

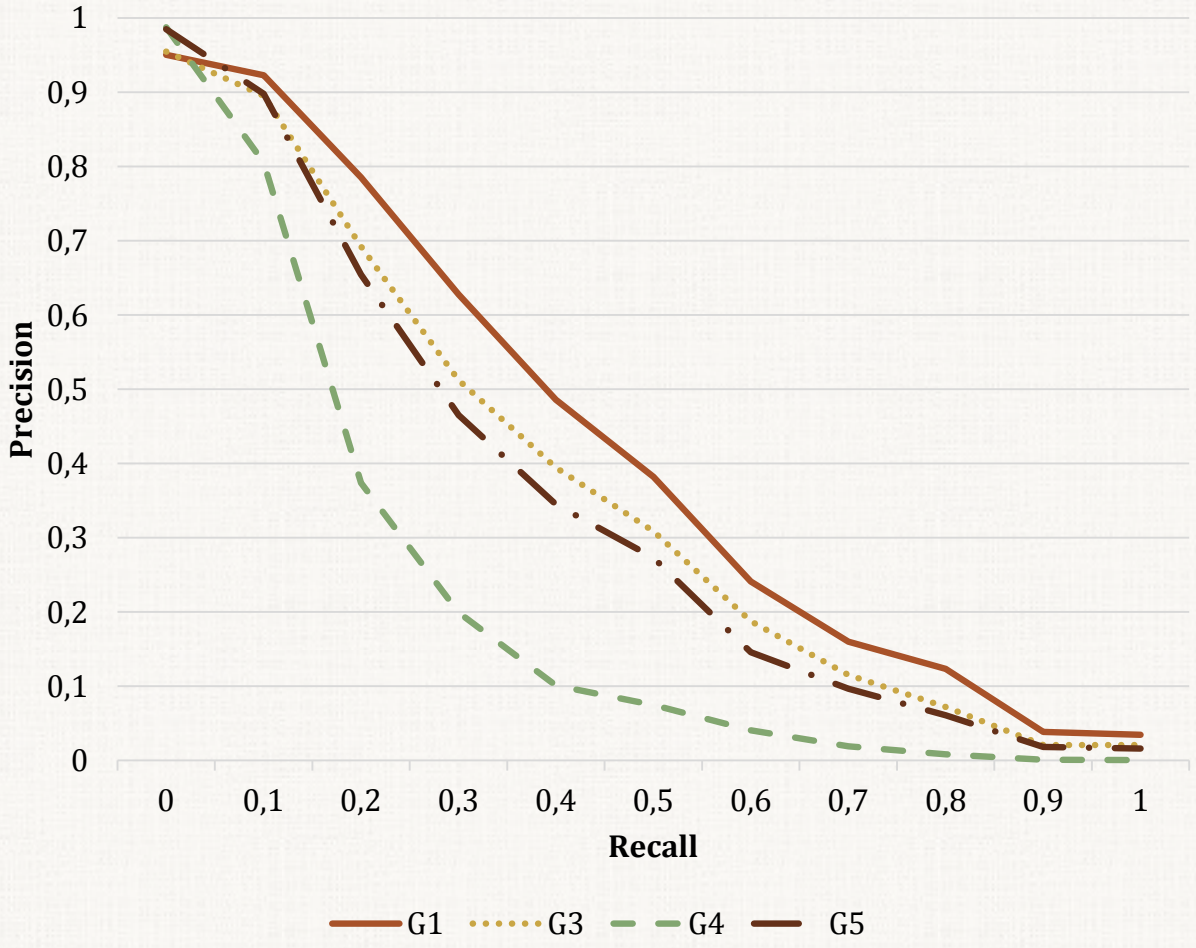
Method	P@5				MAP				NDCG (Binary)				NDCG			
	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.617	0.611	0.599	0.609 (**1**)	0.428	0.419	0.402	0.416 (**1**)	0.653	0.640	0.621	0.638 (**1**)	0.671	0.657	0.640	0.560 (**1**)
G3	0.596	0.568	0.506	0.556 (**2**)	0.397	0.372	0.321	0.363 (**2**)	0.551	0.518	0.457	0.509 (**2**)	0.569	0.536	0.474	0.526 (**2**)
G4	0.351	0.341	0.313	0.335 (**4**)	0.219	0.209	0.187	0.205 (**4**)	0.386	0.363	0.319	0.356 (**4**)	0.400	0.376	0.331	0.369 (**4**)
G5	0.597	0.55	0.477	0.543 (**3**)	0.385	0.347	0.280	0.337(**3**)	0.569	0.513	0.424	0.502 (**3**)	0.586	0.531	0.440	0.519 (**3**)

MODERN DATASET

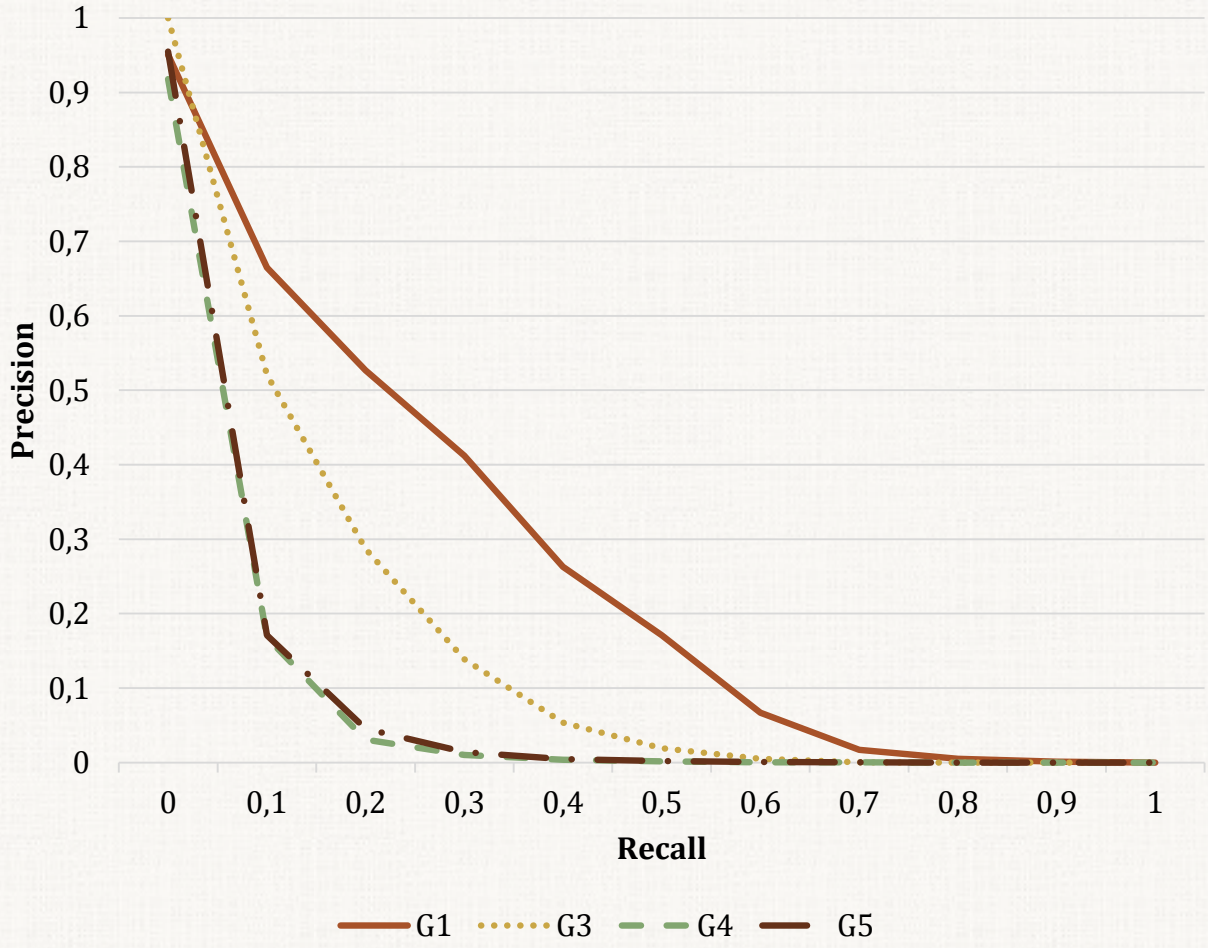
Method	P@5				MAP				NDCG (Binary)				NDCG			
	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average	Overlapping Threshold			Average
	0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8		0.6	0.7	0.8	
G1	0.541	0.541	0.535	0.539 (**1**)	0.265	0.265	0.259	0.263 (**1**)	0.491	0.484	0.473	0.483 (**1**)	0.491	0.485	0.474	0.483 (**1**)
G3	0.429	0.422	0.399	0.417 (**2**)	0.170	0.165	0.152	0.163 (**2**)	0.310	0.301	0.277	0.296 (**2**)	0.310	0.301	0.277	0.296 (**2**)
G4	0.250	0.241	0.211	0.234 (**4**)	0.095	0.089	0.077	0.087 (**4**)	0.218	0.195	0.161	0.191 (**4**)	0.218	0.195	0.161	0.191 (**4**)
G5	0.264	0.247	0.223	0.245 (**3**)	0.100	0.092	0.081	0.091 (**3**)	0.229	0.201	0.168	0.199 (**3**)	0.229	0.202	0.168	0.200 (**3**)

TRACK I: SEGMENTATION-BASED - PRECISION – RECALL CURVES

BENTHAM DATASET



MODERN DATASET



FINAL RANKING

TRACK I: SEGMENTATION-BASED

Rank	Method	Score
1	G2	10
2	G1	14
3	G3	24

TRACK II: SEGMENTATION-FREE

Rank	Method	Score
1	G1	8
2	G3	16
3	G5	24
4	G4	32

AND THE WINNER IS:

FOR TRACK I – SEGMENTATION-BASED:

Method **G2** which has been submitted by **Jon Almazán, Albert Gordo, Ernest Valveny**

affiliated to the:

Computer Vision Center, Universitat Autònoma de Barcelona, Spain.

TRACK II – SEGMENTATION-FREE

Method **G1** which has been submitted by **Alon Kovalchuk, Lior Wolf, Nachum Dershowitz**

affiliated to the:

Blavatnik School of Computer Science, Tel-Aviv University, Israel.